# Lecture 1: Chapters 1-3.2 Intro, Sampling, Surveys

□Variable Types and Roles
□Summarizing Variables
□4 Processes of Statistics
□Data Production; Sampling
□Various Study Designs; Surveys

# Example: *What Statistics Is All About*

- **Background**: Statistics teacher has a large collection of articles and reports of a statistical nature.
- **Question:** How to classify them?
- **Background:** Statistics students are faced with a collection of exam problems at the end of the semester.
- **Question:** How to choose the right procedures to solve them?

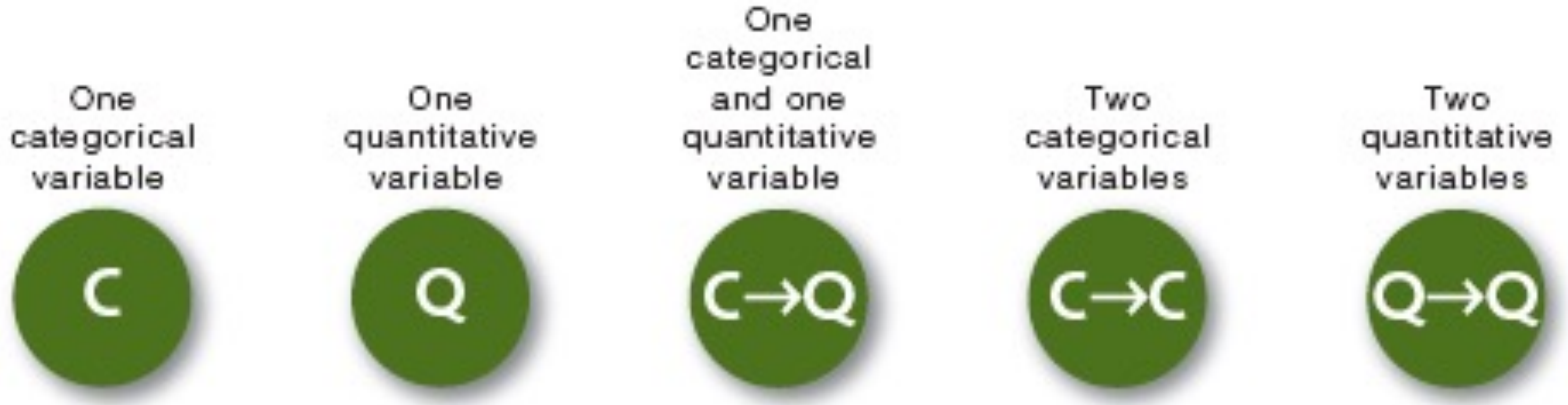
# **Example:** *What Statistics Is All About*

- **Response (to both questions):** Statistics is all about…

***Looking Ahead:*** *Identifying what kind of variables are involved is the key to classifying statistics problems and choosing the right solution tool.*

# The Five Variable Situations

- □ When studying relationships between two variables, we often think of one as explanatory and the other as response.
- □ Depending on the variables' types and roles, we consider five possible situations.


One categorical variable
C
One quantitative variable
Q
One categorical and one quantitative variable
C→Q
Two categorical variables
C→C
Two quantitative variables
Q→Q

# Example: *Identifying Types of Variables*

- **Background**: Consider these headlines…
  - *Dark chocolate might reduce blood pressure*
  - *Half of moms unaware of children having sex*
  - *Vampire bat saliva researched for stroke*
- **Question:** What type of variable(s) does each article involve?
- **Response:**
  - Dark chocolate or not is ________________
    blood pressure is ________________
  - Being aware or not of children having sex is ____________
  - Bat saliva or not is ________________
    stroke recovery is probably _______________

*Practice: 1.2 p.11*

# **Example:** *Categorical Variable Giving Rise to Quantitative Variable*

- **Background:** Individual teenagers were surveyed about drug use.

| Teenager | Marijuana? | Harder Drugs? |
|---|---|---|
| #1 | Yes | Yes |
| #2 | No | No |
| #3 | No | No |
| #4 | Yes | No |
| . . . | . . . | . . . |

- **Question:** What type of variable(s) does this involve?
- **Response:**
  - marijuana or not is ______________
  - harder drugs or not is ______________

*Practice: 1.6a p.12*

# **Example:** *Categorical Variable Giving Rise to Quantitative Variable*

- **Background:** Percentages of teenagers using marijuana or hard drugs are recorded for a sample of countries.

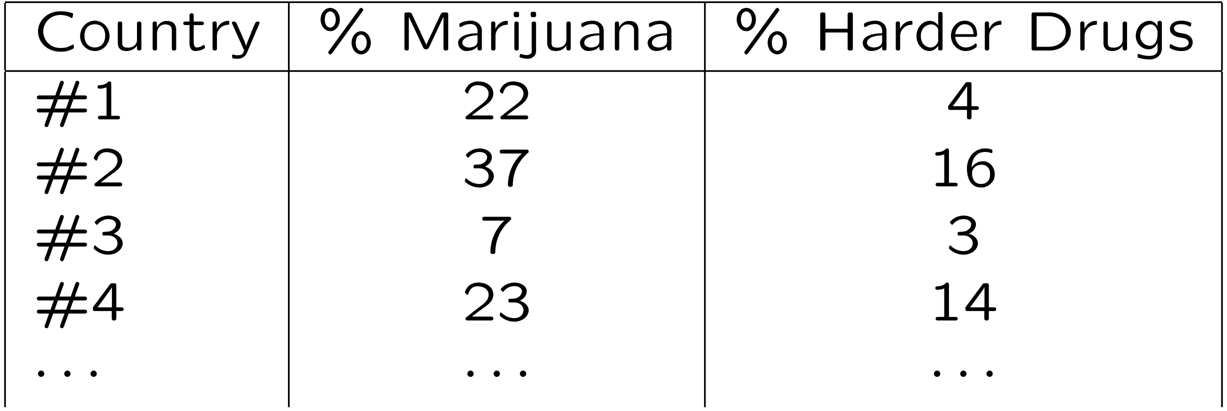
| Country | % Marijuana | % Harder Drugs |
|---|---|---|
| #1 | 22 | 4 |
| #2 | 37 | 16 |
| #3 | 7 | 3 |
| #4 | 23 | 14 |
| . . . | . . . | . . . |

- **Question:** What type of variable(s) does this involve?
- **Response:**
  - percentage using marijuana is ________________
  - percentage using harder drugs is ________________

*Practice: 1.6b p.12*

# **Example:** *Categorical Variable Giving Rise to Quantitative Variable*

- **Background:** Percentages of teenagers using marijuana or hard drugs are recorded for a sample of countries.



| Country | % Marijuana | % Harder Drugs |
|---|---|---|
| #1 | 22 | 4 |
| #2 | 37 | 16 |
| #3 | 7 | 3 |
| #4 | 23 | 14 |
| ... | ... | ... |

- **Question:** What type of variable(s) does this involve?
- **Response:** (another perspective)
  - type of drug (marijuana or harder drugs) is ___________
  - % using the drugs is _____________

# **Example:** *Quantitative Variable Giving Rise to Categorical Variable*

- **Background**: Researchers studied effects of dental X-rays during pregnancy.
  - *First approach:* X-rays or not; baby's weight
  - *Second approach:* X-rays or not; classify baby's wt. as at least 6 lbs. (considered normal) or below 6 lbs.
- **Question:** What type of variable(s) does each approach involve?
- **Response**:
  - X-rays or not is ____________; baby's weight is ____________
  - X-rays or not is ____________;

    baby's wt. at least 6 lbs. or below 6 lbs. is ________________

*Practice: 1.8 p.12*

# Definitions

- **Data**: recorded values of categorical or quantitative variables
- **Statistics:** science concerned with
  - gathering data about a group of individuals
  - displaying and summarizing the data
  - using info from data to draw conclusions about larger group

*(All these skills are essential in both academic and professional settings.)*

# Summarizing Data

- **Categorical** data:
  - **Count:** number of individuals in a category
  - **Proportion:** count in category divided by total number of individuals considered
  - **Percentage:** proportion as decimal × 100%
- **Quantitative** data: **mean** is sum of values divided by total number of values

# **Example:** *Summarizing Variables*

- □ **Background**: Recent research unearthed evidence that for a short period of time, a few women voted in America (specifically, New Jersey) around 1800: "*…In total, the lists include 163 unique women's names, with women casting about 208 of the 2,695 documented votes. Overall, they found, about 7.7% of total votes recorded were cast by women…*"
- □ **Question:** What type of variable is involved, and how is it summarized?
- □ **Response**: gender of voters is ____________, summarized with ________________________________________________

*Hint: think about who or what are the individuals. What information is recorded for each of them?*

*Practice: 1.10 p.12*

# **Example:** *Summarizing Variables*

- **Background**: A 2019 lawsuit alleged inequities in average pay by the software giant Oracle: *"Oracle's…female, Black, and Asian employees with years of experience are paid as much as 25% less than their peers."*
- **Question:** What type of variable is considered, and how is it summarized?
- **Response**: ______________________,

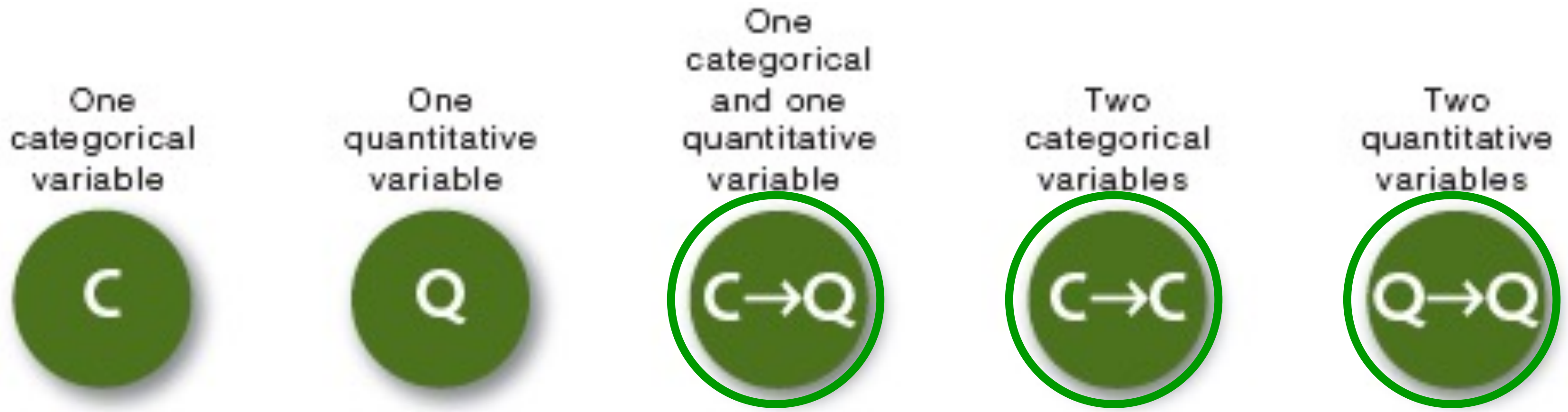
  summarized with ________________

*A Closer Look: When comparing quantitative values for two or more categorical groups, we sometimes quantify the difference by reporting what percentage higher or lower one mean is compared to the other.*

*Practice: 1.11 p.12*

# Roles of Variables

When studying relationships between two variables, we often think of one as explanatory and the other as response.


One categorical variable
C
One quantitative variable
Q
One categorical and one quantitative variable
C→Q
Two categorical variables
C→C
Two quantitative variables
Q→Q

# **Example:** *Identifying Types and Roles*

- **Background:** Consider these headlines---
  - *Men twice as likely as women to be hit by lightning*
  - *Do Oscar winners live longer than less successful peers?*
- **Questions:** What types of variables are involved?

  For relationships, what roles do the variables play?
- **Responses:**
  - Gender is ________________ and _________________

    Hit by lightning or not is ______________ and ________________
  - Winning an Oscar or not is

    _______________ and ________________

    Life span is _______________ and _______________

*Practice: 1.17 p.13*

# **Example:** *More Identifying Types and Roles*

- **Background:** Consider these headlines---
  - *35% of returning troops seek mental health aid*
  - *Smaller, hungrier mice*
  - *Average rent for an apartment in Pittsburgh is $1256 (March 2021)*
- **Questions:** What types of variables are involved?

  For relationships, what roles do the variables play?
- **Responses:**
  - Seeking mental health aid or not is ______________________
  - Size is ____________ and _____________

    Appetite is __________ and ____________
  - Rent is ____________________________

# Definitions

- A **random** occurrence is one that happens by chance alone, and not according to a preference or an attempted influence.
- **Probability:** formal study of the chance of occurring in a random situation.
- **Statistical Inference**: drawing conclusions about population based on sample.

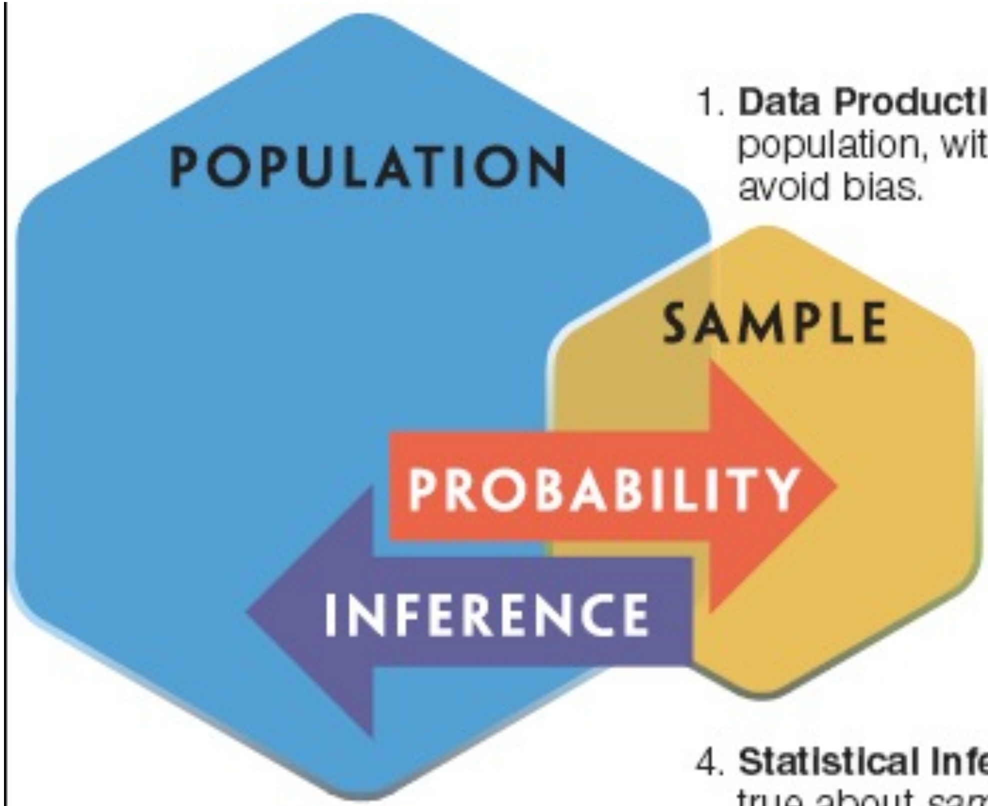
***Looking Ahead:*** *Probability and Inference are linked through their roles in the 4-stage process of Statistics.*

# Statistics as Four-Stage Process

- □ **Data Production**
- □ **Displaying and Summarizing**
- □ **Probability**
- □ **Statistical Inference**

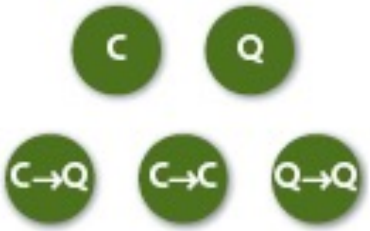
***Looking Ahead:*** *Besides the word "probability", a Probability statement may use the word "chance" or "likelihood" (the only synonyms available).*

# Four Processes of Statistics



1. **Data Production:** Take sample data from the population, with sampling and study designs that avoid bias.

2. **Displaying and Summarizing:** Use appropriate displays and summaries of the sample data, according to variable types and roles.



3. **Probability:** Assume we know what's true for the *population;* how should **random** *samples* behave?

4. **Statistical Inference:** Assume we only know what's true about *sampled* values of a single variable or relationship; what can we **infer** about the larger *population?*

# Data Production

- □ Use a good **sampling design** to get an unbiased sample so we can ultimately generalize from sample to population (Part 4)
- □ Create a good **study design** so what we learn is unbiased summary of what's true about the variables in our sample (Part 2)

# Sampling:  First Step in Data Production

*Each student chooses a whole number at random from 1 to 20.*

*Are the selections truly unbiased?  A show of hands may indicate that certain numbers are favored over others…*

# Definition

- □ **Bias:** tendency of an estimate to deviate in one direction from a true value

*Some sources of bias:*

selection bias: due to unrepresentative sample, rather than to flawed study design

- □ sampling frame doesn't match population
- □ self-selected (volunteer) sample
- □ haphazard sample
- □ convenience sample
- □ non-response

# **Example:** *Bias in Sampling*

□ **Background**: Professor seeks opinions of 5 from 50 class members about textbook…

1. *Have students raise hand if they'd like to give an opinion*
2. *Sample the next 5 students coming to office hours*
3. *Pick 5 names "off the top of his head"*

□ **Questions:** Is each sampling method biased? If so, how?

□ **Responses:**

**1.** ____________________________________________

____________________________________________

**2.** ____________________________________________

____________________________________________

**3.** ____________________________________________

____________________________________________

*Practice: 2.6 p.26*

# **Example:** *More Bias in Sampling*

- □ **Background**: Professor seeks opinions of 5 from 80 class members about textbook…

1. *Assign each student in classroom a number (1, 2, 3, …), then use software to select 5 at random…*
2. *Take a random sample from the roster of students enrolled; mail them anonymous questionnaire…*

- □ **Questions:** Is each sampling method biased? If so, how?
- □ **Responses:**

**1.** ______________________________

______________________________

**2.** ______________________________

______________________________

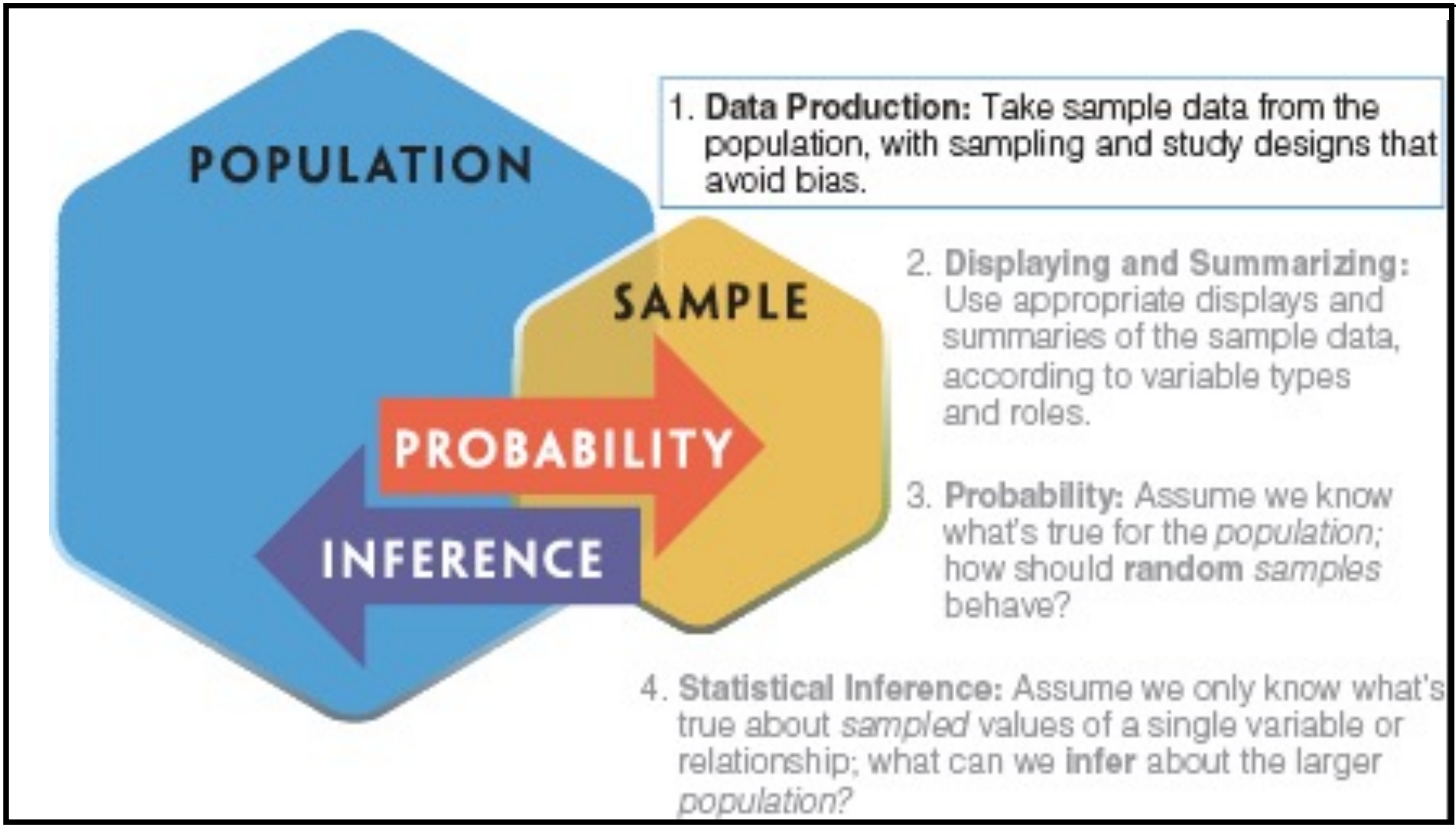
# Definitions

- **Probability sampling plan** incorporates randomness in the selection process so rules of probability apply.
- **Simple random sample** is taken at random and without replacement.
- **Stratified random sample** takes separate random samples from groups of similar individuals (strata) within the population.

# Definitions

- **Cluster sample** selects small groups (clusters) at random from within the population (all units in each cluster included).
- **Multistage sample** stratifies in stages, randomly sampling from groups that are successively more specific.
- **Systematic sampling plan** uses methodical but non-random approach (select individuals at regularly spaced intervals on a list).

# Four Processes of Statistics



1. **Data Production:** Take sample data from the population, with sampling and study designs that avoid bias.
2. **Displaying and Summarizing:** Use appropriate displays and summaries of the sample data, according to variable types and roles.
3. **Probability:** Assume we know what's true for the *population;* how should **random** *samples* behave?
4. **Statistical Inference:** Assume we only know what's true about *sampled* values of a single variable or relationship; what can we **infer** about the larger *population?*

The Data Production stage entails not only selecting a sample, but also designing a study to learn about the variables of interest for that sample.

# Definitions

- **Observational study**: researchers record variables' values as they naturally occur (can be retrospective or prospective).
- **Sample survey:** observational study with self-reported values, often opinions
- **Experiment:** researchers manipulate explanatory variable, observe response
- **Anecdotal evidence:** personal accounts by one or a few individuals selected haphazardly or by convenience. *(To be avoided.)*

# One Possible Study Design: Sample Surveys

- **Types of Study Design**
  - Experiment: researchers control explanatory variable
  - Observational study: values occur naturally
    - Special case: sample surveys (often self-reported).
- **Two steps in Data Production**
  - Obtain an unbiased sample.
  - Assess variables' values to obtain unbiased summary of sample.
    - Design survey questions to assess values without bias.

# Example: *Formulating a Survey Question*

- **Background:** A popular 2005 movie sparked speculation: how common is it for a 40-year-old male to be a virgin?
- **Question:** Assuming you had a representative sample of 40-year-old males, what survey question would you ask to find out what proportion are virgins?

Students can jot down question & discuss after covering issues in survey question design.

# Sample Survey Design: Issues to Consider

□ Open vs. closed questions

□ Unbalanced response options

□ Leading questions or planting ideas with questions

□ Complicated questions

□ Sensitive questions

□ Hard-to-define concepts

# Example: *Open vs. Closed Questions*

- □ **Background:** An exam may feature these…
- □ **Questions:**

1. What kind of question is this?
   (a) open (b) closed
2. What is an open question?

- □ **Responses:**

1. (Choose one) (a) open (b) closed
2. ______________________________________________

*Practice: 3.60 p.66*

# Definitions

- An **open question** does not have a fixed set of response options.
- A **closed question** either provides or implies a fixed set of possible responses.

# Example: *Overly Restrictive Options*

- **Background:** A neuroscientist asked survey respondents, “How often do you dream in color? Answer always/sometimes/never”
- **Question:** What is the most important improvement that should be made to this survey question?
- **Response:**

*Practice: 3.11 p.44*

# **Example:** *Unbalanced Response Options*

- **Background:** 91% of Americans surveyed rated their own health as good to excellent.
- **Questions:**
  - Is this result surprising to you?
  - If so, does it seem unexpectedly high or low?
- **Responses:**
  - ________________________________
  - ________________________________

# **Example:** *Unbalanced Response Options*

- □ **Background:** 91% of Americans surveyed rated their own health as good to excellent. Options provided were

  Excellent / Very Good / Good / Fair / Poor

- □ **Question:** Now is the result surprising?
- □ **Response:**

*Practice: 3.14 p.44*

# Example: *Deliberate Bias*

- **Background:** The following question was posted on www.a-human-right.com: If my child or my spouse were assaulted, I would…(choose one)
  1. Run away and hope my kid or spouse can keep up
  2. Be a good witness so I can tell the cops what happened later
  3. Try to convince the attacker to stop through verbal persuasion
  4. Fight to stop the attack
- **Question:** Do we know what response the surveyor wants us to choose?
- **Response:**

*Practice: 3.16 p.44*

# Deliberate Bias

If it’s clear what response the surveyor wants, then the results are not useful from a statistical standpoint.

# **Example:** *Complicated Question*

- **Background:** A telephone surveyor asked a homemaker to agree or disagree with this:

  "I don't go out of my way to purchase low-fat foods unless they're also low in calories."
- **Question:** How can this survey question be improved?
- **Response:**

*Practice: 3.21d p.45*

# Example: *A Controversial Question*

- **Background:** Anonymous PA Youth Survey given to 6th-12th public school students asked:

  How old were you when you first…

  - got suspended from school
  - got arrested
  - carried a handgun…etc.

  Choose: never have / 10 or younger / 11 / 12 / …/17
- **Questions:**
  - Why did parents object?
  - Why was the question worded this way?
- **Responses:**
  - ______________________________
  - ______________________________

*Practice: 3.25 p.45*

# Example: *Keyboards for Sense of Anonymity*

- □ **Background:** A stats computer tutor was piloted in a class where students consented to be identified by name. Still, one student filled in the text boxes with obscenities.
- □ **Question:** Why did the student write inappropriately in the computer lab, and not on his hard-copy homeworks or exams?
- □ **Response:**

***A Closer Look:*** *This tendency is used to researchers' advantage when seeking responses to sensitive questions.*

*Practice: 3.27 p.46*

# Example: *Hard-to-Define Concepts*

- **Background:** A survey found 19% of Americans believe money can buy happiness.
  - R. Frost: “Happiness makes up in height for what it lacks in length.”
  - A. Camus: “But what is happiness except the simple harmony between a man and the life he leads?”
- **Questions:**
  - By Frost’s definition, can money buy happiness?
  - By Camus’s definition, can money buy happiness?
  - What definition of happiness were respondents using?
- **Responses:**
  - Frost: ____________________
  - Camus: ___________________
  - Respondents: _________________________

*Practice: 3.28 p.46*

# Example: *Formulating a Survey Question*

- **Background:** Earlier we asked, "Assuming you had a representative sample of 40-year-old males, what survey question would you ask to find out what proportion are virgins?"
- **Question:** Are you satisfied with the phrasing of your question; if not, how would you rephrase it?
- **Response:** Consider
  - *Open or closed?*
  - *If closed, what response options are provided?*
  - *Is question designed to elicit honest responses?*
  - *Is the concept well-defined?*

# Lecture Summary *(Introduction, Sampling)*

- **Variables**
  - Categorical or quantitative
  - Explanatory or response
- **Summaries**
  - **Categorical:** count, proportion, percentage
  - **Quantitative:** mean
- **4 Processes:** Data Production, Displaying and Summarizing, Probability, Inference
- **Data Production:** need unbiased sampling and unbiased study design
- **Types of Bias**
- **Types of Samples**

# **Lecture Summary** *(Sample Surveys)*

- □ Open vs closed questions
- □ Unbalanced response options
- □ Leading questions
- □ Complicated questions
- □ Sensitive questions
- □ Hard-to-define concepts